D692–D699 *Nucleic Acids Research, 2013, Vol. 41, Database issue* *Published online 28 November 2012*
doi:10.1093/nar/gks1176

GenColors-based comparative genome databases for small eukaryotic genomes

Marius Felder[1,*], Alessandro Romualdi[2], Andreas Petzold[1], Matthias Platzer[1], Jürgen Sühnel[3] and Gernot Glöckner[4,*]

[1]Genome Analysis Group, Leibniz Institute for Age Research–Fritz Lipmann Institute, Beutenbergstr 11, 07745 Jena, [2]Universitätsklinikum Jena, Erlanger Allee 101, 07747 Jena, [3]Biocomputing Group, Leibniz Institute for Age Research–Fritz Lipmann Institute, Beutenbergstr 11, Jena and [4]Institute for Biochemistry I, Medical Faculty, University of Cologne, Joseph-Stelzmann-Straße 52, D-50931 Köln, Germany and Molecular Biology Group, Leibniz-Institute of Freshwater Ecology and Inland Fisheries, Müggelseedamm 301, 12587 Berlin, Germany

Received August 13, 2012; Revised October 10, 2012; Accepted October 28, 2012

ABSTRACT

Many sequence data repositories can give a quick and easily accessible overview on genomes and their annotations. Less widespread is the possibility to compare related genomes with each other in a common database environment. We have previously described the GenColors database system (http://gencolors.fli-leibniz.de) and its applications to a number of bacterial genomes such as *Borrelia*, *Legionella*, *Leptospira* and *Treponema*. This system has an emphasis on genome comparison. It combines data from related genomes and provides the user with an extensive set of visualization and analysis tools. Eukaryote genomes are normally larger than prokaryote genomes and thus pose additional challenges for such a system. We have, therefore, adapted GenColors to also handle larger datasets of small eukaryotic genomes and to display eukaryotic gene structures. Further recent developments include whole genome views, genome list options and, for bacterial genome browsers, the display of horizontal gene transfer predictions. Two new GenColors-based databases for two fungal species (http://fgb.fli-leibniz.de) and for four social amoebas (http://sacgb.fli-leibniz.de) were set up. Both new resources open up a single entry point for related genomes for the amoebozoa and fungal research communities and other interested users. Comparative genomics approaches are greatly facilitated by these resources.

INTRODUCTION

With the expanding use of new sequencing technologies a growing number of prokaryotic and eukaryotic genomic sequences become available at an ever increasing pace. Sequence databases like GenBank [(1), http://www.ncbi.nlm.nih.gov/genbank/] provide static data repositories for these genomes, from which datasets can be retrieved and further analysed. However, comparative genomics has to be performed primarily outside of these databases if more than simple Basic Local Alignment Search Tool (BLAST) searches are to be performed. This is why, mainly in metazoa, databases that are dedicated to specific aspects of genome analysis were set up. Examples of such resources are Golden Path [(2), http://genome.ucsc.edu] and Ensembl Genomes [(3), http://www.ensemblgenomes.org]. Ensembl includes a number of selected genomes and the user can browse genomes, but only a limited number of tools for comparative analysis are provided. In contrast to Ensembl, Golden Path offers a variety of comparative genomic datasets and the user can visualize additional information like sequence similarities and compile features of a certain genomic region. These databases have proven invaluable for assessing and using reference genomes by the scientific community. With GenColors, we use a complementary approach with special emphasis on the comparison of closely related bacterial or small eukaryote genomes including the option of homology-based gene annotation. We laid special emphasis on the easy setup and access of the databases to be particularly useful for specialized research communities. Thus, the whole system can be transferred, but we also will add more dedicated databases in the future.

The GenColors system we have developed makes it easy to apply comparative genome analysis tools to

*To whom correspondence should be addressed. Tel: +49 3641 656453; Fax: +49 3641 656255; Email: mfelder@fli-leibniz.de
Correspondence may also be addressed to Gernot Glöckner, Tel: +49 221 478 7375; Fax: +49 221 478 6979; Email: gernot.gloeckner@uni-koeln.de

a set of related genomes. Such analyses are, for example, required for a better understanding of species evolution. The built-in automated analysis tools are especially valuable when no or only scarce information for a manual curation of genomes is available (4).

GenColors was originally created to facilitate annotation, analysis and presentation of bacterial genomes (up to 10 Mb) from the earliest to the final stages of a *de novo* low redundancy sequencing genome project (5). The system has been used for the first time for the genomic annotation of *Borrelia garinii* in 2004 (6). Since then GenColors was used for various prokaryotic genome projects comprising *Borrelia*, *Legionella*, *Leptospira* and *Treponema* (7). With the emergence of next-generation sequencing technologies, *de novo* assemblies of more complex and larger genomes of non-model eukaryotes are available. Such genome assemblies are often characterized by a considerable number of gaps recalcitrant to sequencing and/or assembly. Moreover, eukaryote genomes have complex gene structures (introns and exons, alternative splice sites, etc.) and are normally larger than prokaryote genomes and thus pose additional challenges for such a system. By modifying the structure of the MySQL relational database where all data are stored, we now have adapted GenColors to handle also small eukaryotic genomes (up to several 100 Mb). Additionally also eukaryotic gene structures (introns and exons) can now be displayed, and an InterProScan analysis (8) of eukaryotic protein sequences was added. We have used the database and the tools therein for a first pass annotation and comparative analysis of two fungal species [*Arthroderma benhamiae*, *Trichophyton verrucosum*, (9)] and four social amoebas [*Dictyostelium fasciculatum*, *Dictyostelium discoideum*, *Polysphondylium pallidum* and *D. lacteum* (10)]. Further recent developments include whole genome views and genome list options. For bacterial genome browsers, a prediction and display of potential horizontal gene transfer (HGT) (11) was added. All dedicated publicly available databases for prokaryotic and eukaryotic genomes can be accessed via one central entry point, the GenColors web site http://gencolors.fli-leibniz.de. With these databases, we provide easy access to comparative tools together with the annotated sequence data.

The genome comparison tools within GenColors were used to annotate newly sequenced genomes and thus provided a means to a quick overview on these genomes in comparison with reference sequences. Because annotations can be made directly in the database by experts from the community, the GenColors browsers are a versatile alternative to databases with a manual curation team, which normally require extensive resources. In the following, we discuss briefly for the 'Social Amoebas Comparative Genome Browser' (SACGB) the available data and analysis tools. All databases available from http://gencolors.fli-leibniz.de have the same features and thus can readily be used in the same way.

GenColors FEATURES

Examples for usage of GenColors features in SACGB

The available GenColors tools for visualizing, comparing and analyzing the data are organized in three categories: 'General information', 'Search' and 'Genome comparison'. SACGB currently contains the genome sequences of four so-called social amoebae. The database was heavily used to annotate and analyse three of the genomes by comparing them to the high-quality reference genome of *D. discoideum* (12).

General information

There are two entry points to obtain general information about genomes included in the database. Genome statistics can be explored using the 'Genomes' page. More data can be accessed using the 'General information' category of the 'Methods' page. This category includes gene lists and an option to generate lists for all genes of a specific Clusters of Orthologous Genes (COG) functional category across all genomes (13). Different lists of features of one or more genomes can be displayed in one representation, facilitating comparative analysis. Gene lists can be generated for complete genomes and individual genomic elements (contigs/scaffolds, plasmids or chromosomes). The lists usually include information on gene name, locus tag, GenBank description, genomic element, start position, length, strand and G and C nucleotide (GC) content. The genes can be sorted according to all these features. The 'Genome lists' option accessible using the 'Methods' button enables a quick overview of whole genomes, with detailed statistics on each segment or chromosome. Table 1 shows the first 5 entries of such a list for the *D. lacteum* genome. Gene lists with a more compact output are generated by the 'Quick search' and 'Advanced search' tools. In the first case, only information on gene name, locus tag, GenBank description and genomic element are displayed, and in the second case, the gene name together with all best bidirectional hits (BBHs) and the best five UniProt (14) hits are shown. If a specific genome is chosen from the submenu, a list of all annotated Coding Sequences (CDS) will be generated. For all genomes, linear genome plots can be generated, visualizing a variety of genome features like protein-coding, RNA, transfer RNA, ribosomal RNA genes for both strands, repeat regions and GC content (Figure 1). Furthermore, bacterial genomes can be displayed as circular plots. An additional graphical overview of the genomes is offered by the '(Clickable) linear whole genome views' option. Genes of a complete element are colored according to their COG classification and directly linked to the respective 'Gene information sheet' (see below).

On the level of individual genes, the information is provided by 'Gene information sheets'. Figure 2 shows the 'Gene information sheet' of the *D. discoideum* gene *ybl1*. Because this is the reference genome, a link is provided that gives direct access to the manually curated reference database dictyBase (http://dictybase.org) (Figure 2D). Access to these sheets is provided by either links from the graphical genome overview or a link from a

Table 1. Genome list with statistical information for each element. Only the first 5 entries are shown

No.	Genomic element	Genome length (bp)	GC content of genomic element(s) (%)	Number of genes	Average gene length (bp)	Gene density (number of genes per Mb)	GC content (%)	Pyrimidine content of genes (%)	Purine content of genes (%)	Fraction of hypothetical genes (%)	Fraction of putative genes (%)
1	*D. lacteum*, DDLF0006c05.r1	3994	40.13	2	1863	500.75	40.28	51.18	48.81	100	0
2	*D. lacteum*, DDLF0038h09.f4	26 265	28.3	16	1281	609.17	30.03	46.59	53.4	62.5	6.25
3	*D. lacteum*, F4PJNLW01A00V1	1 465 623	30.2	698	1612	476.24	32.48	47.4	52.59	55.44	6.01
4	*D. lacteum*, F4PJNLW01AQJ8S	45 691	28.84	21	1723	459.6	31.12	47.48	52.51	38.09	4.76
5	*D. lacteum*, F4PJNLW01B0TO9	1 030 050	29.62	496	1566	481.53	31.92	47.14	52.85	57.25	6.25

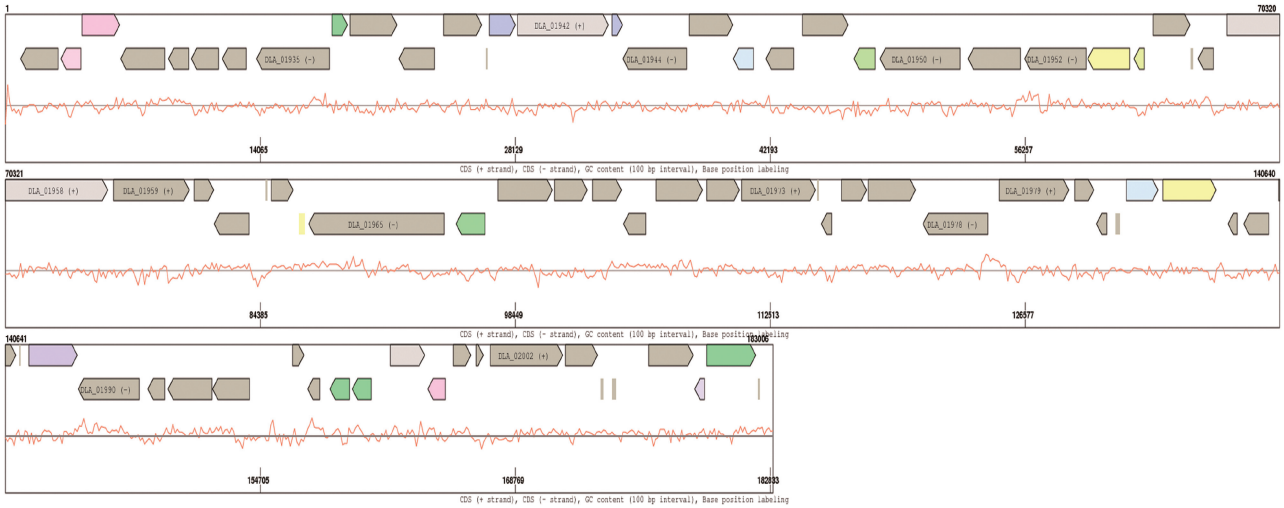

Figure 1. Linear genome plot displaying a 183 kb fragment of the *D. lacteum* genome. Genes on the + and − strands are shown together with the GC content. The coloring is according to COG functional categories.

gene list. These sheets include a gene environment representation, which can be enlarged by the 'Zoom out' button to cover a larger part of the flanking regions of the respective gene (Figure 2A). If a more detailed picture of the gene is required, a click on the 'Show Basepair View' button displays the DNA sequence of both strands of the gene together with their translation. In the central part of the sheet after the gene and corresponding protein sequence more information on the protein is provided (Figure 2B). In case of reference genomes, which are curated, links to external databases like Pfam (15) or Kyoto Encyclopedia of Genes and Genomes (16), etc. are provided for this gene. The data sheets thus provide all information available for the gene in question. For comparative purposes, we provide pre-computed BBHs of all proteins to all other proteins of each genomic element included in the browser database (see genome comparison). Because each genomic element is handled as a unit, a given protein can have more than one BBH, if the second BBH is located on a different element than the first. This feature thus gives a quick overview, whether a protein is present as a single ortholog in a genome or is either part of a gene family (Figure 2E) or fragmented in one of the assemblies. Furthermore, the five best Swiss-Prot/TrEMBL hits as well as related InterPro and Gene Ontology classifications (17) are given for this specific protein. To further improve the analysis of eukaryotic proteins, the results of an InterProScan analysis of these proteins are also displayed on the 'Gene information sheets' (Figure 2C).

Special features in prokaryote browsers

HGT is an important evolutionary mechanism in bacterial genomes and a major factor in a species' adaptation to environmental niches. Therefore, in prokaryotic GenColors genome browsers, HGT predictions were added in the 'Gene information sheets'. The predictions are generated with the program SIGI-HMM (18). In addition to the SIGI-HMM results, related information such as GC content, presence of mobility genes or transfer RNAs is offered. The data can be displayed both in tabular form and as clickable global genome view (Figure 3), allowing a fast genome-wide overview about potentially horizontal transferred genes.

Search options and custom list generation

To explore the increasing amount and complexity of genomic data, the system offers several ways of searching.

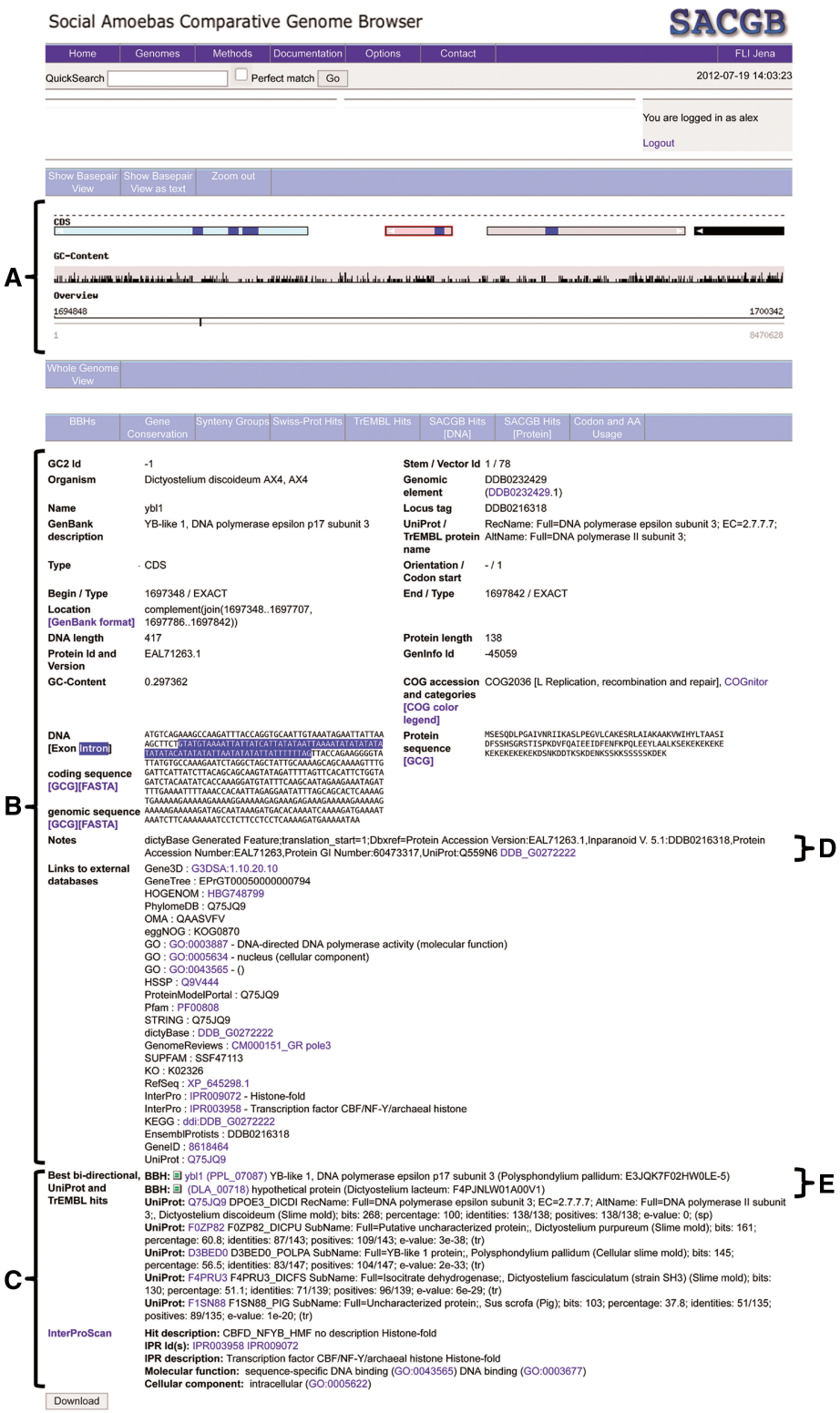

Figure 2. 'Gene information sheet' of the *D. discoideum* gene *ybl1*. (**A**) 'Gene environment graph'. (**B**) Basic gene information, including Swiss-Prot description and all external database links. (**C**) Best-bidirectional hits (BBHs) to all other genomic elements included in the browser database and the five best Swiss-Prot/TrEMBL hits as well as related InterPro and Gene Ontology classifications are indicated. (**D**) Link to dictyBase. (**E**) Links to BBHs and alignment views.

There is a 'Quick search' option on top of each page from which the whole database content can be searched. The result is a list of genes that contains the entered text string either in the gene names, descriptions or locus tags. Under the 'Methods' button, several possibilities for data analysis are accessible. The 'Advanced search' option

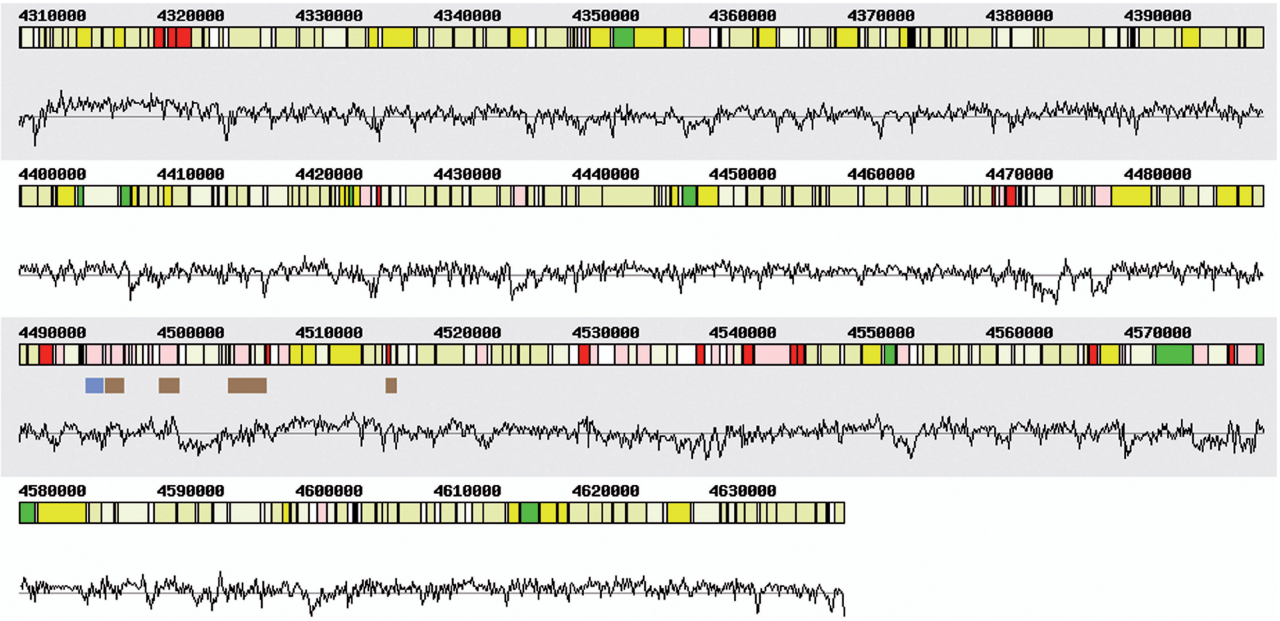

Figure 3. Whole genome view of HGT. Red: HGT genes, type info—blue: integrases, type info—brown: repetitive regions.

allows searching using a combination of up to 22 different data types depending on the data included in the specific browser. These include gene identifiers/description, gene lengths, quality data, genomes or genomic elements, COG categories, PROSITE sequence patterns (19) and the complete external database information provided by Swiss-Prot. For example, this option can be used to identify gene families by using Pfam and InterPro identifiers as search string. Entering the string 'IPR00696' in the InterProScan field of the SACGB database, for example, retrieves 14 entries from 4 genomes containing this domain as a list.

Sequence based searches are done using BLAST. The BLAST server searches the whole sequence space of the GenColors database with a DNA or protein sequence and provides a graphical output of the matching segment. Each genome can also be searched separately. Thus, external users can check their gene of interest against the comparative genome database for presence or absence. The BLAST output contains a link to the 'Gene information sheet' of the matching protein. In the 'Gene information sheet', the pre-computed best BLAST hits against global databases are included to enable direct analyses without the need to search for similarities.

Genome comparison

Orthology, defined as homology through speciation (20), is a crucial concept in evolutionary biology and is essential for comparative genomic studies. In GenColors, orthologous relationships between genes are identified by BLAST BBHs of the corresponding protein sequences. In our database system, at least 30% sequence identity and a length ratio of at least 0.3 between the length of the hit in respect to the length of the protein are required by default to define a BBH. The BBHs between pairs of genomic segments determined by this approach form the basis for a wide range of comparative analyses on protein variation, gene core sets and synteny. A statistical survey of sequence differences between two proteins of a BBH provides information on amino acid differences and on synonymous and non-synonymous base substitutions/exchanges in tabular form. The ratio of non-synonymous to synonymous substitutions in a protein-coding gene may reflect the relative influence of positive/negative selection and neutral evolution and is calculated by the program Syn-SCAN (21) that adopts a method by Nei and Gojobori (22). The protein variation analysis is complemented by the recently improved 'Alignment viewer' (Figure 4), which is accessible through an icon on the 'Gene information sheet' (Figure 2E). All identified protein sequence pairs are globally aligned using the EMBOSS program needle (23). The 'Alignment viewer' displays statistical data and provides the user with a graphical overview of pairwise alignments between protein and DNA sequences, respectively. To further improve the usability of this tool, nucleotides and amino acid residues can be colored to highlight polymorphic sites, indels and specific amino acid patterns.

The BBHs are also used to determine 'Gene core sets', which are defined as groups of genes with BBHs for all possible pairs of organisms in the dataset. These groups may constitute a 'core' of essential genes common to these related organisms. A set of closely related and putatively homologous genes allows a better understanding of the relationships and organization of genomes and provides also an overview on gene family expansions.

The term 'synteny' describes a conserved gene order along the chromosomes of different species (24). In GenColors, synteny groups are defined as two or more neighbouring genes in one genomic element that have neighbouring BBHs in another genomic element of either the same or a different species. Neighbouring genes may be interrupted by up to five genes that have

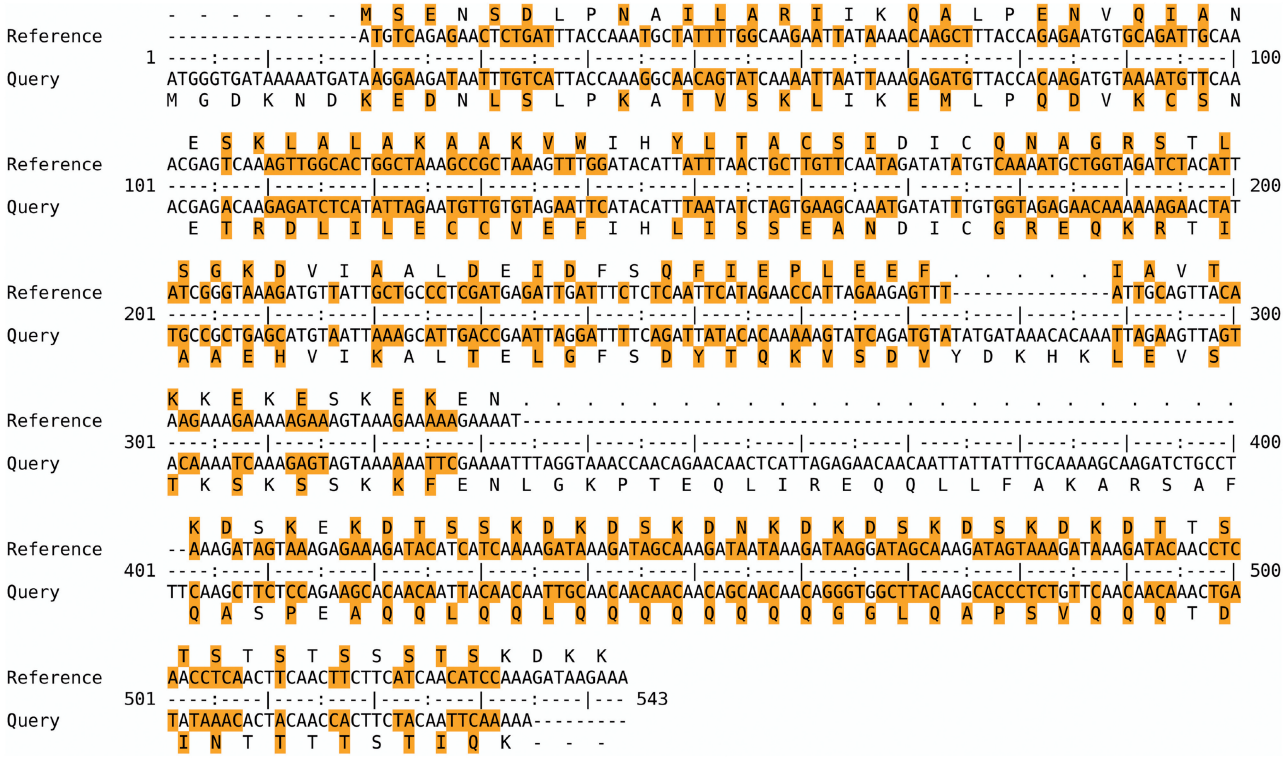

Figure 4. 'Alignment viewer' displaying a combined amino acid and nucleotide alignment.

no BBH in the counterpart genomic element. Genomic region comparisons depicting the structural annotation of syntenic and non-syntenic genes can be displayed as graphical maps and as tables. Additionally there is an option for displaying the gene order within the syntenic gene groups of a synteny group, as illustrated in Figure 5. Finally, the 'Gene conservation' tool indicates for each gene if there are homologous genes present in any genome included in the browser database and if these genes are synteny group members. Taken together, these capabilities enable users to detect gene structures and gene contents present in all species, those specific to a single genome and genes occurring in one clade but absent in another.

INPUT/OUTPUT AND ANNOTATION

The 'Gene information sheets' represent the main starting point for manual gene annotation. Search outputs can be at the same time entry points for subsequent analyses with this subset of genes. Select buttons in each list make it possible to compile and store individual gene lists that can be used for further work and to share them with other users. The most simple way to generate such lists is a keyword search, which then retrieves database entries. From the entry list, one can further select entries and then store it. These lists can then be extended with additional gene lists and used as a subset for further searches and comparisons. The individual lists are accessible if a user is registered. The registration button is accessible through each database front page and is valid for all databases. The registration ensures that a user can have access to his/her previously defined lists. Curators with specific privileges are allowed to edit and change the annotations for each gene as well as to add notes to the datasheets. Thus, this function makes manual curation of the databases possible. Modifications are visible to all users and are individually logged and time stamped in the database. Semi-automated annotation can be performed by an option where in a BBH list the description of a reference gene can easily be transferred to the orthologous gene by simply clicking on a transfer button.

All genomic sequence data and annotations in the databases are available for download from the 'Genomes' section. Annotations can be exported in either GenBank flat file or in Sequin feature table format. The latter files can be directly used for the GenBank submission procedure using Sequin (http://www.ncbi.nlm.nih.gov/Sequin/index.html). Data files can be obtained in bulk and as specific sets for specific genomic elements. Genomic and intergenic sequences can also be displayed and retrieved using the recently added 'Sequence retrieval' option. Specific sequence download coordinates can be defined using this tool and sequences can be downloaded in FASTA and Genetics Computer Group (GCG) format.

BACKEND

The backend of GenColors includes PERL scripts and a MySQL relational database system that stores the results of the manual annotations and pre-computed automated analyses. Additionally, a web server like Apache, BioPerl

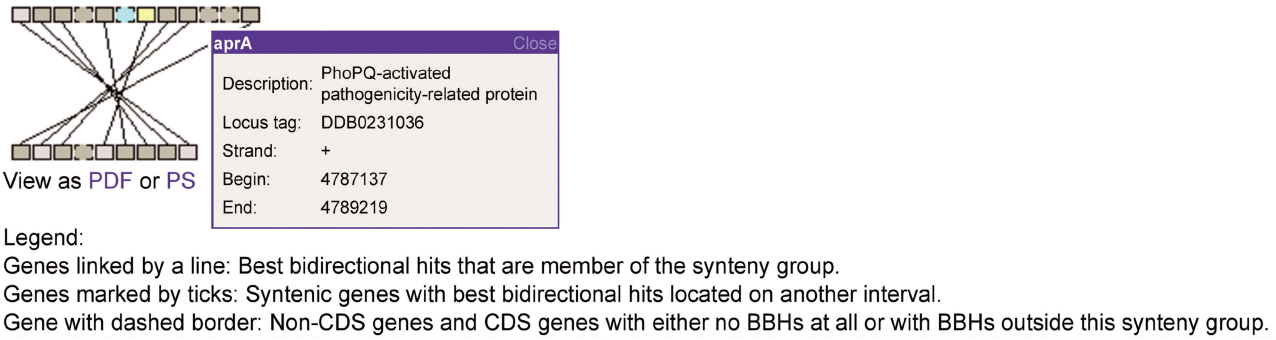

Figure 5. Synteny group with inverted gene order. When moving the mouse pointer over a gene box, description, locus tag, strand and sequence range is shown, a click on the gene box opens the 'Gene information sheet'.

and EMBOSS are required for running a GenColors platform. Thus, the whole system can also easily be ported to other locations and therefore the source code together with a tutorial how to set up the system is also available under the GNU GPL license from the authors for local installation. On request, we also offer to extend existing GenColors databases with additional genomes or to set up new ones for sets of species for users.

CONCLUSIONS AND FUTURE PERSPECTIVES

With the emergence of next-generation sequencing technologies, the amount of high-quality large-scale genomic data increases rapidly. This flood of data has to be reined in with easy-to-use bioinformatics tools. Especially for genomes from species that have a comparable small research community, versatile and comprehensive databases are required. We here provide databases for selected fungi and amoebozoa and some bacteria. We show that using these databases, comparative genome analyses are possible, which otherwise would require the setup of individual analysis pipelines. We will continue to enhance accessibility and visualization of additional features such as gene expression or epigenetics data. A further improvement will be to enable the curation of gene models, i.e. give the possibility to change annotated gene structures.

Taken together, our current set of databases helps to visualize, annotate and analyse data of related organisms, and are invaluable tools as a community resource. Generally, online usage of GenColors-based genome browsers is the preferred mode of application. Yet, the GenColors system also provides a broader platform for comparative genome analysis purposes.

ACKNOWLEDGEMENTS

We thank Kerstin Wagner for her help in setting up and maintaining the SGB external link page as well as in icon design.

FUNDING

Deutsche Forschungsgemeinschaft (DFG; Gl235/1-2). Funding for open access charge: Leibniz Institute for Age Research - Fritz Lipmann Institute.

Conflict of interest statement. None declared.

REFERENCES

1. Benson,D.A., Karsch-Mizrachi,I., Clark,K., Lipman,D.J., Ostell,J. and Sayers,E.W. (2012) GenBank. *Nucleic Acids Res.*, **40**, D48–D53.
2. Dreszer,T.R., Karolchik,D., Zweig,A.S., Hinrichs,A.S., Raney,B.J., Kuhn,R.M., Meyer,L.R., Wong,M., Sloan,C.A., Rosenbloom,K.R. *et al.* (2012) The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Res.*, **40**, D918–D923.
3. Kersey,P.J., Staines,D.M., Lawson,D., Kulesha,E., Derwent,P., Humphrey,J.C., Hughes,D.S., Keenan,S., Kerhornou,A., Koscielny,G. *et al.* (2012) Ensembl Genomes: an integrative resource for genome-scale data from non-vertebrate species. *Nucleic Acids Res.*, **40**, D91–D97.
4. Romualdi,A., Felder,M., Rose,D., Gausmann,U., Schilhabel,M., Glöckner,G., Platzer,M. and Sühnel,J. (2007) GenColors: annotation and comparative genomics of prokaryotes made easy. *Methods Mol. Biol.*, **395**, 75–96.
5. Romualdi,A., Siddiqui,R., Glöckner,G., Lehmann,R. and Sühnel,J. (2005) GenColors: accelerated comparative analysis and annotation of prokaryotic genomes at various stages of completeness. *Bioinformatics*, **21**, 3669–3671.
6. Glöckner,G., Lehmann,R., Romualdi,A., Pradella,S., Schulte-Spechtel,U., Schilhabel,M., Wilske,B., Sühnel,J. and Platzer,M. (2004) Comparative analysis of the *Borrelia garinii* genome. *Nucleic Acids Res.*, **32**, 6038–6046.
7. Glöckner,G., Schulte-Spechtel,U., Schilhabel,M., Felder,M., Sühnel,J., Wilske,B. and Platzer,M. (2006) Comparative genome analysis: selection pressure on the *Borrelia vls* cassettes is essential for infectivity. *BMC Genomics*, **7**, 211.
8. Quevillon,E., Silventoinen,V., Pillai,S., Harte,N., Mulder,N., Apweiler,R. and Lopez,R. (2005) InterProScan: protein domains identifier. *Nucleic Acids Res.*, **33**, W116–W120.
9. Burmester,A., Shelest,E., Glöckner,G., Heddergott,C., Schindler,S., Staib,P., Heidel,A., Felder,M., Petzold,A., Szafranski,K. *et al.* (2011) Comparative and functional genomics provide insights into the pathogenicity of dermatophytic fungi. *Genome Biol.*, **12**, R7.

10. Heidel,A.J., Lawal,H.M., Felder,M., Schilde,C., Helps,N.R., Tunggal,B., Rivero,F., John,U., Schleicher,M., Eichinger,L. *et al.* (2011) Phylogeny-wide analysis of social amoeba genomes highlights ancient origins for complex intercellular communication. *Genome Res.*, **21**, 1882–1891.
11. Gogarten,J.P. and Townsend,J.P. (2005) Horizontal gene transfer, genome innovation and evolution. *Nat. Rev. Microbiol.*, **3**, 679–687.
12. Gaudet,P., Fey,P., Basu,S., Bushmanova,Y.A., Dodson,R., Sheppard,K.A., Just,E.M., Kibbe,W.A. and Chisholm,R.L. (2011) dictyBase update 2011: web 2.0 functionality and the initial steps towards a genome portal for the Amoebozoa. *Nucleic Acids Res.*, **39**, D620–D624.
13. Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
14. The UniProt Consortium. (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **40**, D71–D75.
15. Punta,M., Coggill,P.C., Eberhardt,R.Y., Mistry,J., Tate,J., Boursnell,C., Pang,N., Forslund,K., Ceric,G., Clements,J. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.
16. Kotera,M., Hirakawa,M., Tokimatsu,T., Goto,S. and Kanehisa,M. (2012) The KEGG databases and tools facilitating omics analysis: latest developments involving human diseases and pharmaceuticals. *Methods Mol. Biol.*, **802**, 19–39.
17. The Gene Ontology Consortium. (2012) The Gene Ontology: enhancements for 2011. *Nucleic Acids Res.*, **40**, D559–D564.
18. Waack,S., Keller,O., Asper,R., Brodag,T., Damm,C., Fricke,W.F., Surovcik,K., Meinicke,P. and Merkl,R. (2006) Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. *BMC Bioinformatics*, **7**, 142.
19. Sigrist,C.J., Cerutti,L., de Castro,E., Langendijk-Genevaux,P.S., Bulliard,V., Bairoch,A. and Hulo,N. (2010) PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res.*, **38**, D161–D166.
20. Fitch,W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.*, **19**, 99–113.
21. Gonzales,M.J., Dugan,J.M. and Shafer,R.W. (2002) Synonymous-non-synonymous mutation rates between sequences containing ambiguous nucleotides (Syn-SCAN). *Bioinformatics*, **18**, 886–887.
22. Nei,M. and Gojobori,T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.*, **3**, 418–426.
23. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.
24. Bentley,S.D. and Parkhill,J. (2004) Comparative genomic structure of prokaryotes. *Annu. Rev. Genet.*, **38**, 771–792.